Degrees of Supervision

Darío García García and Robert C. Williamson

1 Australian National University and NICTA

Relations Between Machine Learning Problems
NIPS Workshops
16/12/2011
Outline

1. Introduction
2. General Supervision
3. Generalized Minimum Cross-Entropy
1. Introduction

2. General Supervision

3. Generalized Minimum Cross-Entropy
Lots of “different” learning problems produce the same kind of output but differ on the input information

- Supervised classification
- Supervised classification with noisy labels
- Semi-supervised learning
- Multiple Instance Learning
- Label proportions
- Partial label learning

Can we view all these problems in a common way?
We can think of **clustering** as (transductive) unsupervised classification

- Find the easiest classification problem that can be posed on a given dataset

**What is easy?**

In general, **ambiguity in the labels let us choose our battle**
Standard supervised learning

**Statistical experiment** \( E = (\mathcal{X}, \mathcal{F}, \{P_h\}_{h \in \mathcal{H}}) \)

- **Hypotheses:** \( \mathcal{H} = [H_1, \ldots, H_k] \)
- **Sample space:** \( (\mathcal{X}, \mathcal{F}) \)
- **Probability measures:** \( \{P_h(X)\}_{h \in \mathcal{H}}, X \in \mathcal{F} \)
- **Output:** Decisor \( T : \mathcal{X} \rightarrow \mathcal{H} \)

**Example-based learning:** The probability measures are unknown. Instead, we are given labeled samples

- **Input:** \( \{(x_i, h_i)\}_{i=1}^n, (x_i, h_i) \in \mathcal{X} \times \mathcal{H} \)
  - The examples correspond directly with the desired output
1. Introduction

2. General Supervision

3. Generalized Minimum Cross-Entropy
Compound experiment $E^C = (\mathcal{X}^C, \mathcal{F}^C, \{P_h^C\}_{h \in \mathcal{H}^C})$

- **Compound hypotheses of order $m$:** $\mathcal{H}^C = \mathcal{H}^m$, $h^C = [h_1, \ldots, h_m], h_i \in \mathcal{H}$
- $P_h^C = P_{h_1} \otimes P_{h_2} \otimes \ldots \otimes P_{h_m}$

Introduce two additional spaces

- **Observation space** $(\mathcal{O}, \mathcal{A})$
  - Manifestation of $\mathcal{X}^C$
- **Supervision space** $(\mathcal{S}, \mathcal{B})$
  - Manifestation of $\mathcal{H}$

Recall: the output is still $T : \mathcal{X} \rightarrow \mathcal{H}$
There are stochastic mechanisms (Markov kernels, conditional probabilities, ...) relating the “hidden” and “accessible” spaces

- **Observation model** \( M_O : P_{hC}^C(X) \rightarrow Q_{hC}(O) \)
- **Supervision model** \( M_S : H^C \rightarrow P_s(H^C) \)
  - Standard label: \( P_s(H^C) = \delta(H^C - H_s) \) *fully informative*
  - “Unlabeled”: \( P_s(H^C) = c \) *uninformative*

- **Input:** \( D = \{(o_i, s_i)\}_{i=1}^n, \quad (o_i, s_i) \in O \times S \)

**Recovered experiment:**

\[
\tilde{P}_{hC}^C(X) = \int P(H^C | S) P(S | O) P(O | X^C) dS dO
\]

\[
E^C >> \tilde{E}^C
\]
New challenges

Potential mismatch between the observed space and the space we will use to take decisions
  - Train and test points can live in different spaces
Supervision can be ambiguous
  - The task that we need to solve is not clear. How to choose it?
  - We need an inductive principle
Simple solution: Assume that the problem is as easy as possible
  - How hard is a problem? Loss function
  - Generalized Minimum Cross-Entropy
Particular (but very general) case

$\mathcal{H} = \Delta^k$ (probability estimation)
$\mathcal{O} = 2^X, S = 2^\mathcal{H}$ Aggregation
$S \sim$ set of allowable states

Examples

- Fully supervised: $|o_i| = 1, s_i \in \text{ext}(\Delta^k)$
- Label proportions: $s_i = \frac{1}{n_i} \sum_{x \in o_i} \eta(x)$
- Unsupervised: $s_i = \Delta^k$
Clustering: ML and CML

Two ways of learning a mixture model $P_\theta$ for clustering purposes

\[ \hat{\theta}_{ML} = \arg \max_{\theta} P_\theta(X) = \arg \min_{\theta} \sum_{i=1}^{N} \log \frac{1}{P_\theta(x_i)} \]

Equivalently: Minimize KL divergence

\[ \hat{\theta}_{CML} = \arg \max_{\theta,Y} P_\theta(X|Y) \]

\[ = \arg \min_{\theta,Y} \sum_{i=1}^{N} \underbrace{l_{\log}(y_i, \eta_\theta(x_i))}_{\mathcal{D}_X^{CML}(\eta_\theta)} + \underbrace{\log \frac{1}{P_\theta(x_i)}}_{\mathcal{R}_X^{CML}(P_\theta)} \]

Equivalently: Minimize Cross Entropy

CML finds the simplest regularized classification problem according to log-loss
1 Introduction

2 General Supervision

3 Generalized Minimum Cross-Entropy
Cross Entropy:
How many bits are required to transmit \( x \sim P \) using a code designed for \( Q \)
Influenced by both entropy of \( P \) and “closeness” of \( P \) and \( Q \):

\[
H(P, Q) = H(P) + KL(P \parallel Q)
\]
Generalized Cross Entropy

Shannon information concepts are closely related to the log-loss (compression)

**Generalization:** Substitute for another proper loss (more general statistical problems)

\[ H_\phi(p, \eta) = J_\phi(p) + B_\phi(p, \eta) \]

\( B_\phi: \) Bregman divergence parametrized by convex \( \phi \)

\[ B_\phi(p, \eta) = \phi(p) - \phi(\eta) - \langle p - \eta, \nabla \phi(\eta) \rangle \]

\( J_\phi(p): \) Bayes risk, generalized entropy

\[ J_\phi(p) = \mathbb{E}_{y \sim p}[\phi(y)] - \phi(\bar{p}), \quad \bar{p} = \mathbb{E}[p] \]

Minimize cross-entropy \( \equiv \) Find easy statistical problems which are well approximated by our solution
Proper losses

\[ l_{\phi}(y, \eta) = B_{\phi}(e_y, \eta) = \phi(e_y) - \phi(\eta) - \langle e_y - \eta, \nabla \phi(\eta) \rangle \]

\( e_i \): \(i^{th}\) vertex of the simplex

Point-wise risk

\[ \mathbb{E}_{y \sim p} l_{\phi}(y, \eta) = J_{\phi}(p) + B_{\phi}(p, \eta) \]
Minimize

\[ J_{I_n,\phi}(\eta) = D_{I_n,\phi}(\eta) + R_{I_n}(\eta) \]

\[ I_n = \{s_i, o_i\}_{i=1}^n \]

\( D_{I_n,\phi} : \) Information functional

\( R_{I_n} : \) Data-dependent regularization functional

Information functional

\[ D_{I_n,\phi}(\eta) \propto \sum_i \min_{p \in s_i} H_\phi \left( p, \frac{1}{N_i} \sum_{x_j \in o_i} \eta(x_j) \right). \]
A family of objective functions for learning

\[ J(\eta) = D_{I_n,\phi}(\eta) + R_{I_n}(\eta) \]

\[ I_n = \{s_i, o_i\}_{i=1}^n \]

\( D_{I_n,\phi} \) : Information functional

\( R_{I_n} \) : Data-dependent regularization functional

**Fully Supervised**

\[ D_{I_n,\phi}(\eta) \propto \sum_i H_\phi(s_i, \eta(x_i)) . \]

**Regret**
A family of objective functions for learning

\[ J(\eta) = D_{I_n, \phi}(\eta) + R_{I_n}(\eta) \]

\[ I_n = \{ s_i, o_i \}_{i=1}^n \]

\( D_{I_n, \phi} \): Information functional

\( R_{I_n} \): Data-dependent regularization functional

Unsupervised (with balance penalty)

\[ D_{I_n, \phi}(\eta) \propto \arg \min_{\eta} \sum_i \min_{p \in \text{ext}(\Delta^k)} H_\phi(p, \eta(x_i)) + H_\phi \left( s_G, \frac{1}{N} \sum_i \eta(x_i) \right). \]
Sample applications

Introduce an information functional into a purely smoothness based clustering algorithm

- Classification Spectral Clustering

\[ \eta^* = \arg\min_{\eta} \sum_{i} \min_{p \in \Delta_k} H_\phi (p, \eta_i) + H_\phi \left( s_G, \frac{1}{N} \sum_{i} \eta_i \right) + \tilde{\eta}^T L \tilde{\eta} \]

- Laplacian regularized Label Proportions
Summary

- General view of supervision: Classification, Clustering, Label Proportions, ...
- Ambiguity in the supervision → Choose your own task
- Minimum Generalized Cross Entropy: Find easy problems for a certain loss function
Thanks!